> home > about > feedback > login

US Patent & Trademark Office

Try the *new* Portal design

Give us your opinion after using it.

Citation

Conference on Human Factors in Computing Systems >archive

Proceedings of the SIGCHI conference on Human factors in computing systems >toc

1995 , Denver, Colorado, United States

TileBars: visualization of term distribution information in full text information access

Author
Marti A. Hearst

Sponsor
SIGCHI : ACM Special Interest Group on Computer-Human Interaction

Publisher
ACM Press/Addison-Wesley Publishing Co. New York, NY, USA

Pages: 59 - 66 Series-Proceeding-Article
Year of Publication: 1995
ISBN:0-201-84705-1

> full text > references > citings > index terms > peer to peer

> Discuss > Similar > Review this Article Save to Binder

> BibTex Format

FULL TEXT: Access Rules

html 48 KB

REFERENCES

Note: OCR errors may be found in this Reference List extracted from the full text article. ACM has opted to expose the complete List rather than only correct and linked references.

1 M. Aboud , C. Chrisment , R. Razouk , F. Sedes , C. Soule-Dupuy, Querying a hypertext information retrieval system by the use of classification, Information Processing and Management: an International Journal, v.29 n.3, p.387-396, May-June, 1993

2 Hans C. Arents , Walter F. L. Bogaerts, Concept-based retrieval of hypermedia information: from term indexing to semantic hyperindexing, Information Processing and Management: an

International Journal, v.29 n.3, p.373-386, May-June, 1993

3 Jacques Bertin. (1983) Semiology of Graphics . The University of Wisconsin Press, Madison, WI, 1983. Translated by William J. Berg.

4 Richard Chimera, Value bars: an information visualization and navigation tool for multi-attribute listings, Proceedings of the SIGCHI conference on Human factors in computing systems, p.293-294, May 03-07, 1992, Monterey, California, United States

5 William S. Cooper, Fredric C. Gey, and Aitoa Chen. (1994) Probabilistic retrieval in the TIPSTER collections: An application of staged logistic regression. In Donna Harman, editor, Proceedings of the Second Text Retrieval Conference TREC-2 , pages 57-66. National Institute of standard and Technology Special Publication 500-215, 1994.

6 W. Bruce Croft , Howard R. Turtle, Text retrieval and inference, Text-based intelligent systems: current research and practice in information extraction and retrieval, Lawrence Erlbaum Associates, Inc., Mahwah, NJ, 1992

7 Douglass R. Cutting , David R. Karger , Jan O. Pedersen, Constant interaction-time scatter/gather browsing of very large document collections, Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, p.126-134, June 27-July 01, 1993, Pittsburgh, Pennsylvania, United States

8 Douglass R. Cutting, Jan O. Pedersen, and Per-Kristian Halvorsen. (1991) An object-oriented architecture for text retrieval. In Conference Proceedings of RIAO '91, Intelligent Text and Image Handling , Barcelona, Spain, pages 285-298, April 1991. Also available as Xerox PARC technical report SSL-90-83.

9 Douglass R. Cutting, Jan O. Pedersen, Per-Kristian Halvorsen, and Meg Withgott. (1990) Information theater versus information refinery. In Paul S. Jacobs, editor, AAAI Spring Symposium on Text-based Intelligent Systems , 1990.

10 Dennis E. Egan , Joel R. Remde , Louis M. Gomez , Thomas K. Landauer , Jennifer Eberhardt , Carol C. Lochbaum, Formative design evaluation of superbook, ACM Transactions on Information Systems (TOIS), v.7 n.1, p.30-57, Jan. 1989

11 Edward A. Fox , Matthew B. Koll, Practical enhanced Boolean retrieval: experiences with SMART and SIRE systems, Information Processing and Management: an International Journal, v.24 n.3, p.257-267, May 1988

12 Norbert Fuhr and Chris Buckley. (1993) Optimizing document indexing and search term weighting based on probabilistic models. In Donna Harman, editor, The First Text Retrieval Conference (TREC-1) , pages 89-100. NIST Special Publication 500-207, 1993.

13 Donna Harman, Overview of the first TREC conference, Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval, p.36-47, June 27-July 01, 1993, Pittsburgh, Pennsylvania, United States

14 Martha Alice Hearst, Context and structure in automated full-text information access, University of California at Berkeley, Berkeley, CA, 1995

15 Marti A. Hearst. (1994b) Multi-paragraph segmentation of expository text. In Proceedings of the 32nd Meeting of the Association for Computational Linguistics , June 1994.

16 Marti A. Hearst. (1995) An investigation of term distribution effects in Full-Text Retrieval.

Technical Report Report Number ISTL-QCA-1994-12-06, Xerox PARC, 1995. Submitted for publication.

17 William C. Hill , James D. Hollan , Dave Wroblewski , Tim McCandless, Edit wear and read wear, Proceedings of the SIGCHI conference on Human factors in computing systems, p.3-9, May 03-07, 1992, Monterey, California, United States

18 Brewster Kahle and Art Medlar. (1991) An information system for corporate users: Wide area information servers. Technical Report TMC199, Thinking Machines Corporation, April 1991.

19 Robert R. Korfhage, To see, or not to see— is That the query?, Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval, p.134-141, October 13-16, 1991, Chicago, Illinois, United States

20 S. Kosslyn, S. Pinker, W. Simcox, and L. Parkin. (1983) Understanding Charts and Graphs: A Project in Applied Cognitive Science . National Institute of Education, 1983. ED 1.310/2:238687.

21 J. D. Mackinlay, Automatic design of graphical presentations, Stanford University, Stanford, CA, 1987

22 Alistair Moffat, Ron Sacks-Davis, Ross Wilkinson, and Justin Zobel. (1994) Retrieval of partial documents. In Donna Harman, editor, Proceedings of the Second Text Retrieval Conference TREC-2 , pages 181-190. National Institute of standard and Technology Special Publication 500-215, 1994.

23 Terry Noreault, Michael McGill, and Matthew B. Koll. (1981) A performance evaluation of similarity measures, document term weighting schemes and representations in a Boolean environment. In R. N. Oddy, S. E. Robertson, C. J. van Rijsbergen, and P. W. Williams, editors, Information Retrieval Research , pages 57-76. Butterworths, London, 1981.

24 John Ousterhout. (1991) An X11 toolkit based on the Tcl language. In Proceedings of the Winter 1991 USENIX Conference , pages 105-115, Dallas, TX, 1991.

25 George G. Robertson , Stuart K. Card , Jack D. Mackinlay, Information visualization using 3D interactive animation, Communications of the ACM, v.36 n.4, p.57-71, April 1993

26 Gerard Salton, Automatic text processing: the transformation, analysis, and retrieval of information by computer, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 1989

27 Hikmet Senay and Eve Ignatius. (1990) Rules and principles of scientific data visualization. Technical Report GWU-IIST-90-13, Institute for Information Science and Technology, The George Washington University, 1990.

28 Anselm Spoerri, InfoCrystal: a visual tool for information retrieval & management, Proceedings of the second international conference on Information and knowledge management, p.11-20, November 01-05, 1993, Washington, D.C., United States

29 Edward R. Tufte, The visual display of quantitative information, Graphics Press, Cheshire, CT, 1986

↑ CITINGS *45*

Thomas Tan, Active retrieval results: if a picture is worth a thousand words, how much is a moving picture worth?, CHI '99 extended abstracts on Human factors in computer systems, May 15-20,

1999, Pittsburg, Pennsylvania

David J. Harper , Sara Coulthard , Sun Yixing, A language modelling approach to relevance profiling for document browsing, Proceedings of the second ACM/IEEE-CS joint conference on Digital libraries, July 14-18, 2002, Portland, Oregon, USA

Gareth J. F. Jones , Steven M. Gabb, A visualisation tool for topic tracking analysis and development, Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, August 11-15, 2002, Tampere, Finland

David Modjeska , Vassilios Tzerpos , Petros Faloutsos , Michalis Faloutsos, BIVTECI: a bibliographic visualization tool, Proceedings of the 1996 conference of the Centre for Advanced Studies on Collaborative research, p.28, November 12-14, 1996, Toronto, Ontario, Canada

Ali Hussam , Brian Ford , Jack Hyde , Ali Merayyan , Bill Plummer , Terry Anderson, Semantic highlighting, CHI 98 conference summary on Human factors in computing systems, p.191-192, April 18-23, 1998, Los Angeles, California, United States

Paul Yarin , Hiroshi Ishii, TouchCounters: designing interactive electronic labels for physical containers, CHI '00 extended abstracts on Human factors in computer systems, April 01-06, 2000, The Hague, The Netherlands

Marti A. Hearst , Jan O. Pedersen, Visualizing information retrieval results: a demonstration of the TileBar interface, Conference companion on Human factors in computing systems: common ground, p.394-395, April 13-18, 1996, Vancouver, British Columbia, Canada

Christian Jacquemin , Michèle Jardino, Une interface 3D multi-échelle pour la visualisation et la navigation dans de grands documents XML, Proceedings of the 14th French-speaking conference on Human-computer interaction (Conférence Francophone sur l'Interaction Homme-Machine), p.263-266, November 26-29, 2002, Poitiers, France

Martin Wattenberg , David Millen, Conversation thumbnails for large-scale discussions, CHI '03 extended abstracts on Human factors in computer systems, April 05-10, 2003, Ft. Lauderdale, Florida, USA

Bryce Allen, Information space representation in interactive systems: relationship to spatial abilities, Proceedings of the third ACM conference on Digital libraries, p.1-10, June 23-26, 1998, Pittsburgh, Pennsylvania, United States

Wolfgang Hürst , Gabriela Maass , Rainer Müller , Thomas Ottmann, The "Authoring on the Fly" system for automatic presentation recording, CHI '01 extended abstracts on Human factors in computer systems, March 31-April 05, 2001, Seattle, Washington

Christopher Ahlberg, Cocktailmaps: a space-filling visualization method for complex communicating systems, Proceedings of the workshop on Advanced visual interfaces, May 27-29, 1996, Gubbio, Italy

Aravindan Veerasamy , Nicholas J. Belkin, Evaluation of a tool for visualization of information retrieval results, Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, p.85-92, August 18-22, 1996, Zurich, Switzerland

M. G. Brown , J. T. Foote , G. J. F. Jones , K. Spärck Jones , S. J. Young, Open-vocabulary speech indexing for voice and video mail retrieval, Proceedings of the fourth ACM international conference on Multimedia, p.307-316, November 18-22, 1996, Boston, Massachusetts, United States

MetaSpider: meta-searching and categorization on the Web, Journal of the American Society for Information Science and Technology, v.52 n.13, p.1134-1147, November 2001

Allison Woodruff , Ruth Rosenholtz , Julie B. Morrison , Andrew Faulring , Peter Pirolli, A comparison of the use of text summaries, plain thumbnails, and enhanced thumbnails for Web search tasks, Journal of the American Society for Information Science and Technology, v.53 n.2, p.172-185, January 15, 2002

Michael G. Christel , David B. Winkler , C. Roy Taylor, Multimedia abstractions for a digital video library, Proceedings of the second ACM international conference on Digital libraries, p.21-29, July 23-26, 1997, Philadelphia, Pennsylvania, United States

Paul Yarin , Hiroshi Ishii, TouchCounters: designing interactive electronic labels for physical containers, Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit, p.362-369, May 15-20, 1999, Pittsburgh, Pennsylvania, United States

Emile Morse , Michael Lewis , Kai A. Olsen, Testing visual information retrieval methodologies case study: comparative analysis of textual, icon, graphical, and "spring" displays, Journal of the American Society for Information Science and Technology, v.53 n.1, p.28-40, January 1, 2002

Lucy Terry Nowell , Robert K. France , Deborah Hix , Lenwood S. Heath , Edward A. Fox, Visualizing search results: some alternatives to query-document similarity, Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, p.67-75, August 18-22, 1996, Zurich, Switzerland

Aravindan Veerasamy , Russell Heikes, Effectiveness of a graphical display of retrieval results, ACM SIGIR Forum, v.31 n.SI, p.236-245

Michael Chau , Daniel Zeng , Hinchun Chen, Personalized spiders for web search and analysis, Proceedings of the first ACM/IEEE-CS joint conference on Digital libraries, p.79-87, January 2001, Roanoke, Virginia, United States

Allison Woodruff , Andrew Faulring , Ruth Rosenholtz , Julie Morrsion , Peter Pirolli, Using thumbnails to search the Web, Proceedings of the SIGCHI conference on Human factors in computing systems, p.198-205, March 2001, Seattle, Washington, United States

Staffan Björk , Johan Redström, Window frames as areas for information visualization, Proceedings of the second Nordic conference on Human-computer interaction, October 19-23, 2002, Aarhus, Denmark

Nancy E. Miller , Pak Chung Wong , Mary Brewster , Harlan Foote, TOPIC ISLANDS—a wavelet-based text visualization system, Proceedings of the conference on Visualization '98, p.189-196, October 18-23, 1998, Research Triangle Park, North Carolina, United States

Dietmar Wolfram , Jin Zhang, An investigation of the influence of indexing exhaustivity and term distributions on a document space, Journal of the American Society for Information Science and Technology, v.53 n.11, p.943-952, September 6 2002

Steve Whittaker , Julia Hirschberg , John Choi , Don Hindle , Fernando Pereira , Amit Singhal, SCAN: designing and evaluating user interfaces to support retrieval from speech archives, Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, p.26-33, August 15-19, 1999, Berkeley, California, United States

Ramana Rao , Jan O. Pedersen , Marti A. Hearst , Jock D. Mackinlay , Stuart K. Card , Larry Masinter , Per-Kristian Halvorsen , George C. Robertson, Rich interaction in the digital library, Communications of the ACM, v.38 n.4, p.29-39, April 1995

Catherine Plaisant , Ben Shneiderman , Khoa Doan , Tom Bruns, Interface and data architecture for query preview in networked information systems, ACM Transactions on Information Systems (TOIS), v.17 n.3, p.320-341, July 1999

George W. Furnas , Samuel J. Rauch, Considerations for information environments and the NaviQue workspace, Proceedings of the third ACM conference on Digital libraries, p.79-88, June 23-26, 1998, Pittsburgh, Pennsylvania, United States

Hemant K. Bhargava , Juan Feng, Paid placement strategies for internet search engines, Proceedings of the eleventh international conference on World Wide Web, May 07-11, 2002, Honolulu, Hawaii, USA

Hsinchun Chen , Haiyan Fan , Michael Chau , Daniel Zeng, Testing a cancer meta spider, International Journal of Human-Computer Studies, v.59 n.5, p.755-776, November 2003

Jamey Graham, The reader's helper: a personalized document reading environment, Proceedings of the SIGCHI conference on Human factors in computing systems: the CHI is the limit, p.481-488, May 15-20, 1999, Pittsburgh, Pennsylvania, United States

Min-Yen Kan , Judith L. Klavans, Using librarian techniques in automatic text summarization for information retrieval, Proceedings of the second ACM/IEEE-CS joint conference on Digital libraries, July 14-18, 2002, Portland, Oregon, USA

Harald Reiterer , Gabriela Mußler , Thomas M. Mann , Siegfried Handschuh, INSYDER — an information assistant for business intelligence, Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, p.112-119, July 24-28, 2000, Athens, Greece

Bongwon Suh , Allison Woodruff , Ruth Rosenholtz , Alyssa Glass, Popout prism: adding perceptual principles to overview+detail document interfaces, Proceedings of the SIGCHI conference on Human factors in computing systems: Changing our world, changing ourselves, April 20-25, 2002, Minneapolis, Minnesota, USA

Jonathan Foote , John Boreczhy , Andreas Girgensohn , Lynn Wilcox, An intelligent media browser using automatic multimodal analysis, Proceedings of the sixth ACM international conference on Multimedia, p.375-380, September 13-16, 1998, Bristol, United Kingdom

Marti A. Hearst , Jan O. Pedersen, Reexamining the cluster hypothesis: scatter/gather on retrieval results, Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, p.76-84, August 18-22, 1996, Zurich, Switzerland

Russell C. Swan , James Allan, Aspect windows, 3-D visualizations, and indirect comparisons of information retrieval systems, Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, p.173-181, August 24-28, 1998, Melbourne, Australia

Donald Byrd, A scrollbar-based visualization for document navigation, Proceedings of the fourth ACM conference on Digital libraries, p.122-129, August 11-14, 1999, Berkeley, California, United States

Falk Scholer , Hugh E. Williams, Query association for effective retrieval, Proceedings of the eleventh international conference on Information and knowledge management, November 04-09, 2002, McLean, Virginia, USA

Ricardo Baeza-Yates, Visualization of large answers in text databases, Proceedings of the workshop on Advanced visual interfaces, May 27-29, 1996, Gubbio, Italy

Marti A. Hearst , Chandu Karadi, Cat-a-Cone: an interactive interface for specifying searches and viewing retrieval results using a large category hierarchy, ACM SIGIR Forum, v.31 n.SI, p.246-255

Mei Kobayashi , Koichi Takeda, Information retrieval on the web, ACM Computing Surveys (CSUR), v.32 n.2, p.144-173, June 2000

Michael Berthold , David J. Hand, References, Intelligent data analysis, Springer-Verlag New York, Inc., New York, NY, 2003

↑ INDEX TERMS

Primary Classification:
H. Information Systems
↪ **H.3** INFORMATION STORAGE AND RETRIEVAL
↪ **H.3.3** Information Search and Retrieval
↪ **Subjects:** Retrieval models

Additional Classification:
H. Information Systems
↪ **H.3** INFORMATION STORAGE AND RETRIEVAL
↪ **H.3.3** Information Search and Retrieval
↪ **Subjects:** Query formulation
↪ **H.5** INFORMATION INTERFACES AND PRESENTATION (I.7)

J. Computer Applications
↪ **J.1** ADMINISTRATIVE DATA PROCESSING
↪ **Subjects:** Law

General Terms:
Design, Human Factors

↑ Peer to Peer - Readers of this Article have also read:

◆ Data structures for quadtree approximation and compression
Communications of the ACM 28, 9
Hanan Samet

◆ The state of the art in automating usability evaluation of user interfaces
ACM Computing Surveys (CSUR) 33, 4
Melody Y. Ivory , Marti A Hearst

◆ A lifecycle process for the effective reuse of commercial off-the-shelf (COTS) software
Proceedings of the 1999 symposium on Software reusability
Christine L. Braun

- A recovery algorithm for a high-performance memory-resident database system
  ACM SIGMOD Record 16, 3
  Tobin J. Lehman , Michael J. Carey

- A catalog of techniques for resolving packaging mismatch
  Proceedings of the 1999 symposium on Software reusability
  Robert DeLine

Figure 1: Visualization of results. The highlighted vertical column corresponds to document ranked 14. The title of document ranked 14 document will also be highlighted in the title display window. Clicking the highlighted vertical column brings up the full text of that document.

played. For example, we can expect greater accuracy with a first stage display that shows document titles, authors and subject keywords compared to one that shows just the document titles. When this additional document content is displayed in textual form, the increased accuracy may however bring along a negative effect on perusal time (increase in perusal time). This is because more time is consumed perusing the additional content.

A possible means to addressing this problem of displaying more information in the first stage without increasing perusal effort and perusal time is to display information in some form that does not require as much perusal time and screen space as text. Graphical displays (*visualizations*) of the characteristics of documents which are significant in supporting the decision to peruse or not, could enable set-at-a-time perusal of documents, rather than document-at-a-time perusal of text displays.

In the remainder of this paper, we describe a visualization tool meant to address this issue; describe and present the results of an experiment evaluating the tool; and draw some conclusions about its effectiveness as a first stage display.

2 Visualization tool

The visualization tool is an add-on to a basic interface for an IR system. There is a query window. The titles and ranks of retrieved documents (first stage of display) is shown below the query window. Figure 1 shows the visualization tool corresponding to the query "How has affirmative-action affected the construction-industry, construction projects and public works".

The visualization consists of a series of vertical columns of bars. There is one column of bars for each document. The left-most vertical column corresponds to the document ranked 1 and the right-most vertical column corresponds to the document ranked 150. In each vertical column there are multiple bars - one each for each query word. The height of the bar at the intersection of a query-word-row and a document-column corresponds to the weight of that query word in that document. Moving the mouse cursor over the vertical columns highlights the column directly beneath the mouse cursor and simultaneously highlights the title corresponding to that document in the title-display window. The visualization window is scrollable, in case the number of query words exceeds the available vertical space. The words in the visualization are also stopped and stemmed. Thus the combination of the visualization tool and the title display forms the first stage of display in our system. The basic interface, and the visualization tool utilize the INQUERY retrieval engine, version 2.1p3 [CCH92].

2.1 Response to the need for a concise display of document content

In the Introduction, we discussed the need for a concise first stage display which can also be perused quickly. We believe this visualization scheme to qualify for such a first stage display. It provides information valuable in deciding the relevance of document such as the weight of query concepts in the retrieved documents. The information is also displayed in a highly condensed way, and allows many document surrogates to be perused at one time. Textual display of document surrogates force the user to peruse them a document-at-a-time. However, with this visualization one can infer global patterns such as the following. Suppose we are faced with a search topic where a query term 'q' is so important that all relevant documents will have that query word. We would then ask the following questions: To identify relevant documents, we might ask "Which documents have the important query word 'q'?". To evaluate the goodness of the query, we might ask "Does the important query word 'q' appear in most of the retrieved documents?". When comparing the contribution of two query words, one might ask questions such as "What is the contribution of query word q2 compared to q5?". Answers for such questions seem to emerge from the visualization quickly. Such global perception of data is not possible with text displays that emphasize the *parts* rather than the *whole*. We refer to this kind of global perception as "set-at-a-time perusal", since the information gained is about a set of documents.

The presence or absence of specific significant words can be quickly seen, and it is possible, in one glance, to identify sequences of documents which do, or do not have important contributions from specific query words. For the example search topic ("How has affirmative action affected the construction industry'?"), there are two facets that are central: "affirmative action" and "construction industry". From the visualization tool, we can immediately see that most of the documents are concerned with the "construction industry" and only a portion of them have the term "affirmative action". We can also see that the "affirmative action" concept is spread sparsely throughout the top 70 documents. The graphical format of presentation has some important advantages in that it is more condensed and can be more easily and quickly perused than an equivalent text display.

Effectiveness of a graphical display of retrieval results

Aravindan Veerasamy
College of Computing, 801, Atlantic Drive
Georgia Institute of Technology
Atlanta, Georgia 30332-0280
Email: veerasam@cc.gatech.edu

Russell Heikes
Statistics Center
School of Industrial Systems and Engineering
Georgia Institute of Technology
Atlanta, Georgia 30332
Email: russell.heikes@isye.gatech.edu

Abstract

We present the design of a visualization tool that graphically displays the strength of query concepts in the retrieved documents. Graphically displaying document surrogate information enables set-at-a-time perusal of documents, rather than document-at-a-time perusal of textual displays. By providing additional relevance information about the retrieved documents, the tool aids the user in accurately identifying relevant documents. Results of an experiment evaluating the tool shows that when users have the tool they are able to identify relevant documents in a shorter period of time than without the tool, and with increased accuracy. We have evidence to believe that appropriately designed graphical displays can enable users to better interact with the system.

1 Introduction

The overall concern of all components of an IR system is to present the user as much relevant information as possible. While there has been a lot of work on effective algorithms for retrieving and ranking relevant documents, not much attention has been paid to study the effectiveness of user interface components of IR systems. Apart from retrieval mechanisms, interactive IR systems must also be concerned with the design of appropriate display mechanisms that present the retrieved information in the "best possible manner". We discuss what constitutes "best possible" display by examining a typical user interaction with an IR system. A typical interaction with current IR systems proceeds as follows:

- User in an Anomalous State of Knowledge [BOB82] expresses his information need as a query that is interpretable by the system.
- The system matches the query with the stored documents and retrieves a set of documents. In the case of ranked output systems, the result is ranked in the decreasing order of relevance. Boolean systems may rank the documents in a chronological order.
- At the *first stage* of display, a set of document surrogates for the retrieved documents are displayed to the user. These surrogates typically consist of a combination of titles, author, source, date of publication, etc.
- The user inspects the document surrogates and requests more information (such as the full text if available) about those that look relevant. This leads to a *second stage* of display that provides as much information about the document (in many cases, the complete document itself) as is available in the system.
- After going through a sufficient number of documents, the user quits the session or reformulates the query to retrieve a better set of documents.

In this scheme, the first stage display of document surrogates is meant to provide a concise and accurate indication of document content. The second stage display of documents provides more information about the document. In cases where the document full text may not be available for the second stage (such as a typical online library catalog), users proceed to a third stage where they examine a paper-copy in library bookshelves where the complete document may be available.

Thus as the user progresses from the initial to the later stages of display, that which is displayed is more complete and informative, allowing increasingly accurate relevance judgments. However, since more information is displayed about a document in later stages of display, they are also more time-consuming to peruse. Furthermore, requesting second stage of display may be more costly since some systems charge a certain fee to deliver the full text of documents. Apart from the human frustration of waiting for the delivery of full text, one may have to pay for it monetarily since certain systems charge the user based on connect-time and the volume of downloaded data. Therefore, it is advantageous for the searcher to be reasonably certain about the relevance of a document before requesting a second stage of display.

For the user to make accurate relevance judgments based on the first stage display, the *form* and *content* of first stage of display should provide good indication of what document is about. The *form* of the first stage display should be such that it is quickly perusable – the purpose of the first stage display (of providing a quick and concise indication of document content) is lost otherwise. The *content* of the first stage display should be such that users can make accurate judgments about document relevance.

We can expect an improvement in the accuracy of relevance judgment if more content from the documents are dis-

SIGIR 97 Philadelphia PA, USA
Copyright 1997 ACM 0-89791-836-3/97/7..$3.50

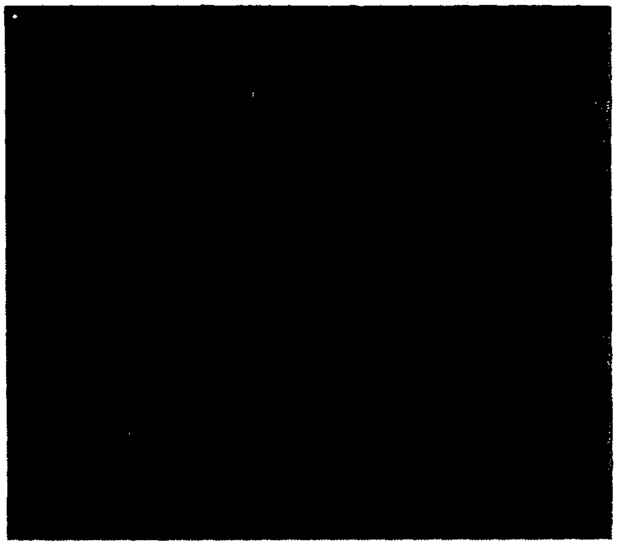

Figure 1: Visualization of results. The highlighted vertical column corresponds to document ranked 14. The title of document ranked 14 document will also be highlighted in the title display window. Clicking the highlighted vertical column brings up the full text of that document.

played. For example, we can expect greater accuracy with a first stage display that shows document titles, authors and subject keywords compared to one that shows just the document titles. When this additional document content is displayed in textual form, the increased accuracy may however bring along a negative effect on perusal time (increase in perusal time). This is because more time is consumed perusing the additional content.

A possible means to addressing this problem of displaying more information in the first stage without increasing perusal effort and perusal time is to display information in some form that does not require as much perusal time and screen space as text. Graphical displays (*visualizations*) of the characteristics of documents which are significant in supporting the decision to peruse or not, could enable set-at-a-time perusal of documents, rather than document-at-a-time perusal of text displays.

In the remainder of this paper, we describe a visualization tool meant to address this issue; describe and present the results of an experiment evaluating the tool; and draw some conclusions about its effectiveness as a first stage display.

2 Visualization tool

The visualization tool is an add-on to a basic interface for an IR system. There is a query window. The titles and ranks of retrieved documents (first stage of display) is shown below the query window. Figure 1 shows the visualization tool corresponding to the query "How has affirmative-action affected the construction-industry, construction projects and public works".

The visualization consists of a series of vertical columns of bars. There is one column of bars for each document. The left-most vertical column corresponds to the document ranked 1 and the right-most vertical column corresponds to the document ranked 150. In each vertical column there are multiple bars - one each for each query word. The height of the bar at the intersection of a query-word-row and a document-column corresponds to the weight of that query word in that document. Moving the mouse cursor over the vertical columns highlights the column directly beneath the mouse cursor and simultaneously highlights the title corresponding to that document in the title-display window. The visualization window is scrollable, in case the number of query words exceeds the available vertical space. The words in the visualization are also stopped and stemmed. Thus the combination of the visualization tool and the title display forms the first stage of display in our system. The basic interface, and the visualization tool utilize the INQUERY retrieval engine, version 2.1p3 [CCH92].

2.1 Response to the need for a concise display of document content

In the Introduction, we discussed the need for a concise first stage display which can also be perused quickly. We believe this visualization scheme to qualify for such a first stage display. It provides information valuable in deciding the relevance of document such as the weight of query concepts in the retrieved documents. The information is also displayed in a highly condensed way, and allows many document surrogates to be perused at one time. Textual display of document surrogates force the user to peruse them a document-at-a-time. However, with this visualization one can infer global patterns such as the following. Suppose we are faced with a search topic where a query term 'q' is so important that all relevant documents will have that query word. We would then ask the following questions: To identify relevant documents, we might ask "Which documents have the important query word 'q'?". To evaluate the goodness of the query, we might ask "Does the important query word 'q' appear in most of the retrieved documents?". When comparing the contribution of two query words, one might ask questions such as "What is the contribution of query word q2 compared to q5?". Answers for such questions seem to emerge from the visualization quickly. Such global perception of data is not possible with text displays that emphasize the *parts* rather than the *whole*. We refer to this kind of global perception as "set-at-a-time perusal", since the information gained is about a set of documents.

The presence or absence of specific significant words can be quickly seen, and it is possible, in one glance, to identify sequences of documents which do, or do not have important contributions from specific query words. For the example search topic ("How has affirmative action affected the construction industry'?"), there are two facets that are central: "affirmative action" and "construction industry". From the visualization tool, we can immediately see that most of the documents are concerned with the "construction industry" and only a portion of them have the term "affirmative action". We can also see that the "affirmative action" concept is spread sparsely throughout the top 70 documents. The graphical format of presentation has some important advantages in that it is more condensed and can be more easily and quickly perused than an equivalent text display.

3 Related work

A number of visualization schemes for information retrieval have been proposed [CRM91, MFH95, Kor91, Spo94, HKW94, ACRS93, AB93] But most of these do not address either the display of query results or the problem of support of relevance assessment. An exception is TileBars [Hea95], but there are some important ways in which TileBars differs from the visualization proposed here.

- TileBars provide information on how the different query facets overlap in different sections of a long document. Our visualization scheme does not provide information at that fine levels of granularity.
- To make the best use of such additional information in TileBars, the user has to decompose the information need into more-or-less orthogonal facets of a query. However, in our visualization, the user can type in the information need as a free-form textual query.
- TileBars presents the document surrogates in a list, making it more difficult than in our tool to gain an overall picture of the query word distribution for a whole set of documents in one glance.
- TileBars seems best suited for long documents, while our visualization scheme seems to be equally effective for short and long documents.

There are a handful of studies that have investigated the effectiveness of document surrogates as content-indicators to enable human relevance judgments [Jan91, Sar69, RRS61, Tho73, MKB78]. None of them studied the effectiveness of graphical displays (visualizations) of document surrogates as content indicators. A result common to all of these studies is that "accuracy" in relevance judgments increases with increasing information (e.g. Title < Abstract < Full text). On the whole, we find that there has been a lack of studies to evaluate the effectiveness of graphical displays of document surrogates as indicators of relevance. This is mainly due to the fact that only recently has it been technologically and economically feasible to render such displays in real-time by the computer. Our study is an attempt to fill that gap.

4 Experimental Setup

In this section, we discuss an experiment to test the effectiveness of the visualization tool as a first stage display, and as a tool to aid effective query reformulation. The part on query reformulation will be discussed in a subsequent paper. We used a portion of the TREC [Har96] database consisting of all of disk1 and disk2 except the "Federal Register" documents. We did not use the Federal Register documents because a high proportion of them did not have a title. We used INQUERY 2.1p3 as the search engine [CCH92]. The retrieval mechanism of the search engine is based on bayesian inference networks using the word occurrence statistics in documents. All of the TREC information topics that we used were very detailed in their description of information need. We picked ten information topics for this study. The criterion used to pick the topics will be discussed below.

A slightly modified version of the *Description* field (mainly removing the introductory words such as "Document will report") was submitted to the retrieval system. 120 documents from the top 150 retrieved documents were obtained and split into two groups as follows: High precision group consisting of 60 documents ranked 1 through 60 and a low precision group consisting of 60 documents ranked 91 through 150. We controlled for precision[1] as a factor in the experiment since we felt that precision might impact the perusal time: Users might more quickly identify non-relevant documents, than the relevant documents. Earlier studies [Sar69, RRS61, MKB78] indicate that precision also influences the ability to judge non-relevance.

Each of the two precision groups were further split into two groups: documents with odd ranks and the documents with even ranks. Thus, there were 4 groups of 30 documents for each information topic: High_precision_even_ranks, High_precision_odd_ranks, Low_precision_even_ranks and Low_precision_odd_ranks. The criterion used to pick the information topics for this study was that the "description" field when used as the query statement must retrieve a set of documents that had a distinct split in the precision values between the high precision group (ranks 1 through 60) and the low precision group (ranks 90 through 150). Since we did not want any overlap in precision values between the high precision group and the low precision group for all the ten chosen topics, we discarded the documents ranked 61 through 90. The precision values in the high precision group for all the chosen topics ranged from 0.43 to 0.6 while those of the low precision group ranged from 0.03 to 0.23.

The experiment we describe was aimed at investigating the effect of visualization on two problems for users:

- accurately identifying relevant documents
- effectively reformulating queries

In this paper, we report on results relevant to only the first of these, but because both problems were addressed in the same experimental design, we describe the entire experiment.

In the experiment, users were given two different types of tasks:

- Task of judging relevance: The users were given the information topic and the search statement used to retrieve documents. They were asked to judge the relevance of each of the 30 documents that were displayed to them as one of
  - relevant to the information topic.
  - non-relevant to the information topic.
  - Unsure.

  For the purposes of the current experiment, clicking the left mouse-button over a document title in the title-display window or over a vertical column in the visualization window marks the document as relevant. Clicking the right mouse button over the title (or the column in the visualization window) marks the document as non-relevant. Middle-clicking it marks the document as "Unsure". Also, left-clicking a query word in the visualization window marks all documents containing that query word as relevant. Right-clicking a query word marks all documents that do not contain that word as non-relevant. Full text or any other information about the documents was not made available to users.
- Query reformulation task: Here the users were asked to "modify the preconstructed query into a form that will retrieve more relevant documents". For half of

[1] *Precision* is the density of relevant documents

the topics, users had the visualization tool and for the other half users did not have the visualization tool – making it a within-subjects, between-topics study.

For the "relevance judgment" task, precision (two levels: high and low) and visualization (two levels: with or without) were controlled in this within-subjects, within-topics study. The even ranked document group was shown with the visualization tool and the odd ranked document group was shown without the visualization tool. The users were not told that the 4 different document groups had two different precision levels. Instead, they were told that the query was issued against 4 different databases and the top 30 documents from each database was presented to them as 4 separate tasks – two with and the other two without visualization. For a given topic, the first task was always a "relevance judgment" task with a high-precision group. The next task was a query reformulation task. The third, fourth and fifth tasks were relevance judgment tasks for the other three groups of 30 documents. The first task was always a relevance judgment task because we wanted the users to have a good feel for the retrieved set of documents before they embarked on the query reformulation task. The first task of relevance judgment was always done with a high-precision document group because, in the real-world the users almost always inspect the top-ranked high-precision document range before they go down the ranks to inspect the low-precision range. Each user did the 5 tasks (4 relevance judgment tasks for the 4 document groups, and one query reformulation task) for 6 information topics, and finally did the search reformulation task for 4 more topics. The 6 topics for which the users did both the relevance judgment and query reformulation were:

- Topic 77: Document will report a poaching method used against a certain type of wildlife.
- Topic 115: Document will report specific consequence(s) of the U.S.'s Immigration Reform and Control Act of 1986.
- Topic 134: Document will report on the objectives, processes, and organization of the human genome project.
- Topic 136: Document will report on attempts by Pacific Telesis to diversify beyond its basic business of providing local telephone service.
- Topic 145: Document will describe how, and how effectively, the so-called "pro-Israel lobby" operates in the United States.
- Topic 197: Document will discuss legal tort reform (a civil wrong for which the injured party seeks a judgment) with regard to placing limitations on monetary compensation to plaintiffs.

The order in which the six topics were presented were balanced across the 37 subjects. The order in which the two visualization conditions appeared for a given topic were also balanced. The order in which the two precision groups appeared in a given topic was not balanced due to the constraint that a high precision group is always the first condition.

The human subjects in this experiment were Georgia Tech undergraduate students enrolled in a one-credit hour class on library searching. Students who participated in the study got full scores in two homework assignments. The complete experiment was split over two days. Subjects were asked to sign a consent form upon arrival. They were then given a demo of the system by the experimenter. They then had a hands-on tutorial where they practiced both the "relevance judgment" task and the "query reformulation" task. Then, they did the 5 tasks for each of the three information topics marking the end of the experiment for the first day. On the second day, they did the 5 tasks for each of the other 3 topics, followed by the "query reformulation" task for 4 other topics.

The subjects were given monetary incentive to do well in the experiment. They were evaluated as follows: We knew a-priori, the relevance of all the documents as given by the TREC assessors. For the relevance judgment task, for each document the user obtained a +1 point if their relevance judgment matches the TREC assessor's judgment, a -1 point if their judgment does not match, and 0 points if they are "Unsure". The user has to judge all of the 30 displayed documents. Thus, for the 4 groups of 30 documents, for the 6 topics, each subject made a total of 4x30x6 = 720 judgments.

		TREC judgment	
		Rel	Not_rel
	Rel	RuRt	RuNt
User judgment	Not_rel	NuRt	NuNt
	Unsure	UuRt	UuNt

The time taken by the subject to complete a task was also noted down. The top 10 quickest subjects with the most points were given monetary awards as follows: All participants were ranked on increasing order of time and decreasing order of points scored. Each participant's rank on both the categories (time and points) were added to get the sum-rank. The participant with the lowest sum rank was considered the best performer. Hence, to do well, one must be both accurate and quick. The top performer was given $50, the second and third performers were given $30 each, the fourth through sixth performers were given $20 each and the seventh through the tenth performers were given $10 each. The participants were told of the rating scheme, so we can assume that they optimized for time and accuracy equally.

Since we claim that graphical display of additional document surrogates does not increase perusal time significantly (due to the set-at-a-time perusal of documents), we predict that the time taken to complete the task for the visualization group will not be significantly higher than the non-visualization group. We also predict an increase in accuracy of relevance judgments for the visualization group, because we claim that very pertinent document surrogate information (i.e., the weight of query words in the retrieved documents) is being displayed in addition to the standard text surrogates such as title and source.

Effectiveness of the visualization tool was measured by what the subjects optimized upon: time, accuracy and the combined time_accuracy rank, where accuracy is the number of correct judgments minus the number of incorrect judgments after discarding the Unsure judgments, i.e., Accuracy = RuRt+NuNt-RuNt-NuRt. However, since the accuracy measure includes the correct judgments, Type I errors and Type II errors all in one score, we split the accuracy measure into distinct components. Here we borrow the analogs of two traditional IR measures "recall" and "precision" and extend them to the interactive situation. In the traditional recall and precision measures, the number of documents that the system judges to be relevant is artificially determined by a cut-off point of top 'X' documents. Let RsRt be the number of documents judged relevant by the system and relevant by the TREC assessor (the user with the original information need). Let RsNt be the number of documents judged relevant by the system and non-relevant by the TREC assessor. Let NsRt be the number of documents judged non-relevant by the system and relevant by the TREC assessor and. Let NsNt be the number of documents judged non-relevant by the system and non-relevant by the TREC assessor.

While traditional "Recall" refers to the ratio of truly relevant documents that the system judged as relevant (i.e., RsRt/(RsRt + NsRt)), we define "Interactive Recall" as the ratio of the truly relevant documents that were judged as relevant by the user (i.e., Interactive Recall = RuRt/(RuRt + NuRt + UuRt)). While traditional "Precision" refers to the ratio of documents judges as relevant by the system that were truly relevant (i.e., RsRt/(RsRt + RsNt)), we define "Interactive Precision" as the ratio of the documents judged as relevant by the user that were truly relevant (Interactive precision = RuRt/(RuRt + RuNt)). Here, a "truly relevant" document is a document that was judged relevant by the TREC assessor. Thus, if we are trying to build an effective first stage display mechanism, we would strive for a display mechanism which would enable a user to pick (and read the full-text of) all of the relevant documents and only the relevant documents displayed. When a user picks a non-relevant document as relevant, it would be time and money wasted perusing a non-relevant document. As a corollary, not being able to pick a relevant document, would be a missing out on relevant information.

However, "Unsure" documents pose a problem. It can be handled in two ways: If we assume that a user always reads the full text of an Unsure document, we should treat the Unsure documents as being judged relevant by the user. Conversely, if a user always skips over an Unsure document, we should treat the Unsure document as being judged non-relevant by the user. Below, we present the analysis with both the interpretations. Thus, if we assume the user to inspect the Unsure documents, we treat the Unsure documents as relevant.
Interactive Recall = (RuRt + UuRt) / (RuRt + NuRt + UuRt)
Interactive Precision = (RuRt + UuRt) / (RuRt + UuRt + RuNt + UuNt)
If we assume the user to not inspect the Unsure documents, we treat the Unsure documents as not-relevant,
Interactive Recall = RuRt / (RuRt + NuRt + UuRt)
Interactive Precision = RuRt / (RuRt + RuNt)

In summary, our hypotheses are:

- Visualization will not increase the time taken to complete the relevance judgment task.
- Visualization will improve the Accuracy of relevance judgments.
- Visualization will improve Interactive Recall.
- Visualization will improve Interactive Precision.

5 Results

Statistical analysis of the experimental data empirically shows that our hypotheses about the relevance judgment task are valid. Since there were 37 subjects, and all subjects did 6 topics with 4 tasks (for each of the 4 groups within the topic) per topic, there were a total of 37 x 6 x 4 = 888 observations. The approach used in all analyses was to construct a least squares, linear additive model of each performance measure as a function of the main effects and interactions of the manipulated experimental variables.

The need for consideration of possible learning/ordering effects, due to the same subjects providing multiple responses at various experimental conditions, is minimized by the balancing of the order in which different experimental conditions are presented to the subjects. However, due to the requirement that within a topic, the high precision condition always be presented first, this balance could not be achieved for this factor. To account for this, the model included a term representing the observation order within subject/topic combination. The design thus allows for independent estimation of all effects except precision and observation order. The analysis presented will focus on the statistical significance of each term assuming the presence of the the other term in the model (i.e on the adjusted sums of squares in the Analysis of Variance (ANOVA) tables), as this provides evaluation of the marginal effect.

The residuals of the models constructed were analyzed to assure reasonable compliance with the normality, independence and constant variance assumptions required for validity of ANOVA,

For the dependent variable "time", the residuals indicated a higher variance for conditions resulting in larger values of time, and hence we transformed time values into $log_{10}(time\ in\ seconds)$ to check for statistical significance. The ANOVA tables for $log_{10}(time)$, accuracy and final score are shown in Tables 1, 2 and 3 respectively. The means and standard errors are shown in table 4. As can be seen from the tables, viz is significantly better than noviz for logtime, accuracy and final score. It is also clear that low precision condition does significantly better than high precision for logtime, accuracy and final score. The interaction effects of precision and visualization are shown in figures 2, 3 and 4 with a 95% confidence interval around the means. When precision is high, visualization does not significantly affect logtime, but when precision is low, there is a decrease in logtime of 0.08. This corresponds to a reduction of 17.2 seconds, nearly a 20% decrease in average time required. Thus we can conclude that the visualization tool helps users in identifying document relevance *more quickly*. It is also interesting to note (from Table 1) that the interaction effect of topic with visualization was not statistically significant, although the main effect of topic was significant. Thus, visualization helps improve speed of judgment irrespective of topic.

For the accuracy measure, there is no significant interaction between precision and visualization as shown by the almost-parallel lines in figure 3. Precision has a huge impact on accuracy, again consistent with previous studies [Sar69, MKB78]. While the effect of visualization on accuracy is significant, it is not as huge as the effect of precision. Users can identify document relevance *more accurately* with the visualization tool than without. The ability of users to identify non-relevant documents as non-relevant is much higher than their ability to identify relevant documents as relevant. This is reflected in the significantly very high accuracy value for low precision than for high precision. It is also interesting to note that (from Table 2) the interaction between topic and visualization was statistically significant. However, the main effect of visualization was much greater than the topic*viz interaction effect.

Final score is a rank measure, which reflects the users ability to *accurately and quickly* identify document relevance. It is plotted in figure 4. Lower values are better for final score. As with accuracy, precision has a much higher impact than visualization, but both variables have a significant effect. Visualization tool improves Final Score and so does low precision. There is a higher proportion of non-relevant documents in the low precision condition. This implies that users can more quickly and accurately judge a non-relevant document as non-relevant compared to judging a relevant document. It is also interesting to note (from Table 3) that the interaction between topic and visualization was statistically significant. However, the main effect of visualization was much greater than the topic*viz interaction effect.

Table 1: ANOVA for log10(time in seconds).

Source	DF	Adj SS	Adj MS	F	P
topic	5	1.59993	0.31999	21.28	0.000
precis	1	0.45954	0.45954	30.56	0.000
viz	1	0.36761	0.36761	24.44	0.000
precis*viz	1	0.29215	0.29215	19.43	0.000
topic*viz	5	0.15612	0.03122	2.08	0.067

Table 2: ANOVA for Accuracy.

Source	DF	Adj SS	Adj MS	F	P
topic	5	10566.13	2113.23	95.04	0.000
precis	1	11842.00	11842.00	532.55	0.000
viz	1	490.54	490.54	22.06	0.000
precis*viz	1	24.67	24.67	1.11	0.293
topic*viz	5	1248.65	249.73	11.23	0.000

Table 3: ANOVA for Final Score.

Source	DF	Adj SS	Adj MS	F	P
topic	5	7187688	1437538	90.98	0.000
precis	1	6789177	6789177	429.68	0.000
viz	1	362841	362841	22.96	0.000
precis*viz	1	133133	133133	8.43	0.004
topic*viz	5	352669	70534	4.46	0.001

As discussed before, accuracy combines the following four items into one: ability to judge relevant and non-relevant documents (RuRt + NuNt), type I error, i.e., wrongly rejecting relevant documents, and type II error, i.e., wrongly accepting non-relevant documents. We feel that identifying non-relevant documents (NuNt) in and of itself is not as important as the other 3 items. For, it is important

- to minimize Type I errors, or else one runs the risk of missing out too many relevant documents..
- to minimize type II errors, or else one runs the risk of wasting too much money and effort in examining non-relevant documents.

We can capture all the interesting data with interactive recall and interactive precision as described in the previous section. In our tables, when users are assumed to treat unsure documents as relevant, the interactive precision and interactive recall are denoted by "iprecwu" and "irecwu" respectively. Correspondingly, when unsure documents are assumed to be treated as non-relevant, interactive precision

Table 4: Least Square Means and Standard errors for Logtime, Accuracy and Final score

Precis	Viz	Logtime	Accur	FinScor
Low	Without	2.01	15.72	353.2
Low	With	1.93	17.54	288.4
High	Without	2.04	5.72	576.1
High	With	2.04	6.87	560.2
STD ERR OF EST		0.009	0.35	9.4

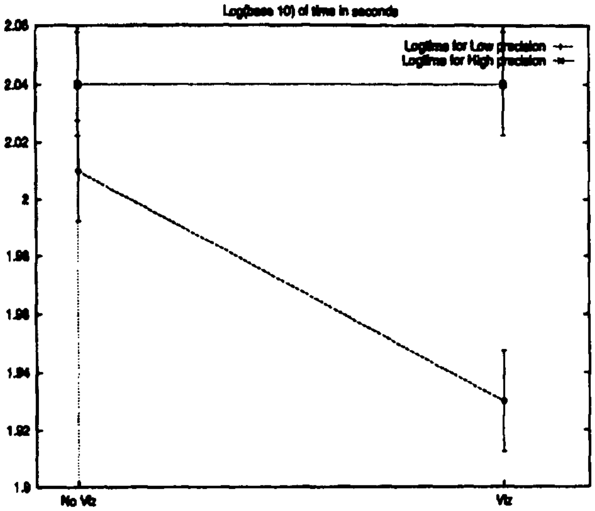

Figure 2: Interaction effects of precision and visualization on logtime.

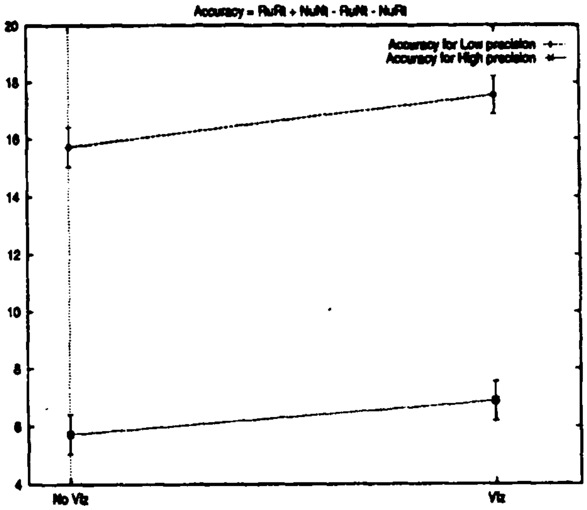

Figure 3: Interaction effects of precision and visualization on accuracy.

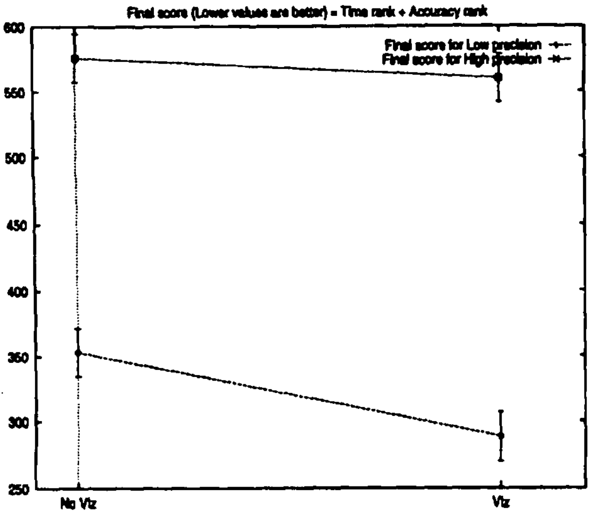

Figure 4: Interaction effects of precision and visualization on Final Score.

and interactive recall are denoted by the mnemonics "iprecwou" and "irecwou" respectively.

In considering the interactive precision measure there are a large number of cases where the values result in responses of zero divided by zero when users did not pick any of the displayed documents as relevant. Rather than eliminate these cases, the raw data (i.e., RuRt, RuNt, NuRt, NuNt, UuRt, UuNt) was aggregated over high and low precision levels for the same viz condition and the interactive precision and interactive recall measures then computed. Thus, for example, for topic 77, the RuRt values for the high_precision_viz case for subject 1 was added to the RuRt value of the low_precision_viz case of the same subject 1 and same topic 77. Now we end up with 444 observations instead of the original 888 observations. This eliminated the need for the "precision" term in the model, although the variability due to this factor is included in the error term. One of the terms is labeled "topic+ord" because the "topic" term also includes some "condition order" effects since for different topics, the four conditions appeared in different orders. The design is now orthogonal to the remaining factors. However for interactive precision when unsure documents are considered non-relevant (iprecwou), there remain 2 cases where the response variable is still zero divided by zero. The result is a design where estimated effects are minimally dependent. Also, there are some quantization errors introduced in the interactive precision measure due to the denominator value being too close to zero[2]. The statistical significance of visualization for Interactive precision and interactive recall (with unsure documents treated as relevant and non-relevant) are shown in tables 5, 6, 7 and 8, and table 9 shows the estimated means.

Visualization had no significant effect on interactive precision when Unsure documents were treated as non-relevant (iprecwou) at the 0.05 level, however, it was significant when Unsure documents were treated as relevant (iprecwu) (See figure 5). Although statistically significant, the absolute increase in interactive precision is very minimal (about 0.015). However, visualization had a significant effect on interactive recall (both when unsure documents were treated as non-relevant (irecwou) and when unsure documents were treated as relevant (irecwu)). Also, in the absolute sense, the improvement in interactive recall due to visualization is approximately 0.07 +/- 0.02 (about a 15% increase). Clearly this is of sufficient magnitude to be of practical importance.

Table 5: ANOVA for Interactive Precision "iprecwou" (Unsure documents treated as non-relevant)

Source	DF	Adj SS	Adj MS	F	P
topic+ord	5	8.15469	1.63094	163.81	0.000
viz	1	0.03065	0.03065	3.08	0.081
topic+ord*viz	5	1.70775	0.34155	34.31	0.000

Table 6: ANOVA for Interactive Precision "iprecwu" (Unsure documents treated as relevant)

Source	DF	Adj SS	Adj MS	F	P
topic+ord	5	6.90892	1.38178	180.62	0.000
viz	1	0.04166	0.04166	5.45	0.021
topic+ord*viz	5	1.02194	0.20439	26.72	0.000

Table 7: ANOVA for Interactive Recall "irecwou" (Unsure documents treated as non-relevant)

Source	DF	Adj SS	Adj MS	F	P
topic+ord	5	3.04486	0.60897	34.92	0.000
viz	1	0.62601	0.62601	35.89	0.000
topic+ord*viz	5	0.72200	0.14440	8.28	0.000

6 Conclusions

We have presented a visualization tool designed to be an effective first stage display of retrieved documents. The results about the query reformulation task and a detailed analysis of all the experimental factors can be found in the thesis by Veerasamy [Vee97]. User experiments empirically show that when precision is low, the visualization tool helps users in identifying document relevance **quicker** by about 20%. Our hypothesis was that the time taken to judge relevance would not be higher for visualization because we claimed that graphically displaying additional information would not take additional time to peruse by enabling set-at-a-time perusal. While this argument is certainly validated by the experimental results, we however see that visualization seems to decrease the time taken. We see only one explanation to this: Users consult visualization before they consult the titles, thereby not looking at the titles of those documents which are clearly non-relevant. Thus they save

[2] For interactive precision when unsure documents are considered non-relevant (iprecwou), there were 2 cases where the denominator had a value of 1, 5 cases of value 2, 6 cases of value 3. For interactive precision when unsure documents are considered relevant (iprecwu), there were 0 cases of denominator values 0 and 3, 1 case of values 1 and 2. Given that there were 444 observation points, these quantization errors are not expected to distort the results much.

Table 8: ANOVA for Interactive Recall "irecwu" (Unsure documents treated as relevant)

Source	DF	Adj SS	Adj MS	F	P
topic+ord	5	2.35410	0.47082	30.21	0.000
viz	1	0.42787	0.42787	27.46	0.000
topic+ord*viz	5	0.41879	0.08376	5.37	0.000

Table 9: Least squares means of iprecwou, iprecwu, irecwou, irecwu

viz	iprecwou	iprecwu	irecwou	irecwu
Without	0.6117	0.5753	0.4454	0.5484
With	0.6284	0.5947	0.5209	0.6108
Std error	0.007	0.006	0.009	0.008

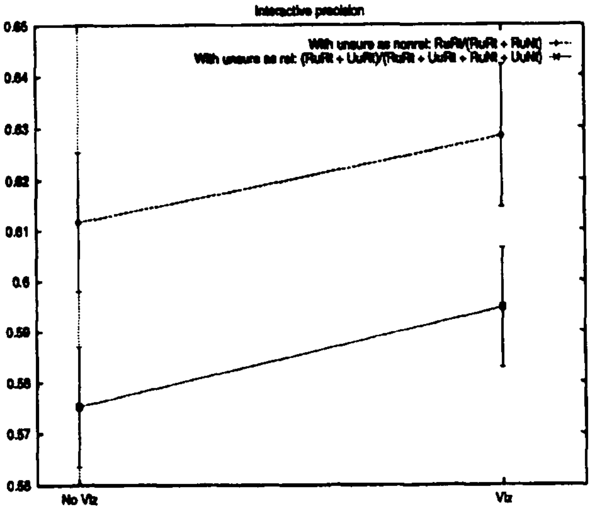

Figure 5: Effect of visualization on interactive precision (when Unsure documents are treated as relevant and non-relevant documents).

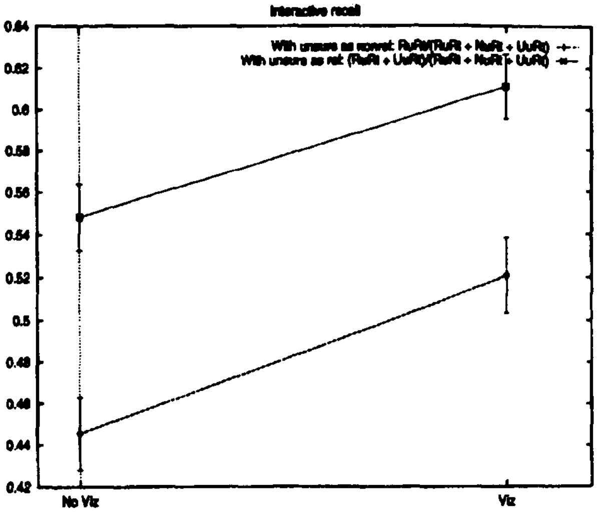

Figure 6: Effect of visualization on interactive recall (when Unsure documents are treated as relevant and non-relevant documents).

the time needed to read titles for those non-relevant documents. This is in agreement with the study by Saracevic [Sar69] which shows that minimal information is needed to say that a document is non-relevant. However, to say that a document is relevant, much more information is needed. This is also confirmed by the fact that the magnitude of time-decrease due to visualization is much higher in the low precision condition than in the high precision condition. On the whole we see confirmation of our argument about set-at-a-time perusal of documents in graphical displays.

The experiment also shows that users with the visualization tool did significantly better in accurate (both in terms of the aggregate "Accuracy" measure and in terms of the broken down measure of "Interactive Recall") identification of document relevance. The result about the influence of precision over relevance judgment Accuracy is in agreement with previous studies by Saracevic [Sar69], and Marcus et.al. [MKB78]. Their studies, like ours, also show that users are better able to judge non-relevance than relevance. However we do not see an interaction between precision and visualization on Accuracy. Thus visualization seems to help increase Accuracy to the same extent irrespective of the density of relevant documents. There is a marked difference in a user's ability to judge the relevance of relevant documents and non-relevant documents. Given this difference, we feel that precision (i.e., the density of relevant documents among the displayed documents) should be a variable that must be controlled in experiments that measure a user's ability to judge relevance. Further, care should be taken in making claims purely based on a compound measure such as "Accuracy" that combines both the ability to correctly identify relevant documents and the ability to correctly identify non-relevant documents.

We broke down the accuracy measure into two components: interactive precision and interactive recall to gain a better understanding of the relevance judgment process. While the effect of visualization tool was marginally significant for interactive precision, it was highly significant for interactive recall. Thus, we can safely say that the visualiza-

Visualization of Large Answers in Text Databases

Ricardo Baeza-Yates*
Dept. of Computer Science, Univ. of Chile
Blanco Encalada 2120, Santiago, Chile
E-mail: rbaeza@dcc.uchile.cl

ABSTRACT

Current user interfaces of full text retrieval systems do not help in the process of filtering the result of a query, usually very large. We address this problem and we propose a visual interface to handle the result of a query, based on a hybrid model for text. This graphical user interface provides several visual representations of the answer and its elements (queries, documents, and text), easing the analysis and the filtering process.

Keywords: visual browsing, visual text databases, visual tools, visual representations, visual query languages, set visualization, visual analysis.

1 Introduction

Full text retrieval systems are a popular way of providing support for on-line text. Their advantage is that they avoid the complicated and expensive process of semantic indexing. From the end-user point of view, full text searching of on-line documents is appealing because a valid query is just any word or sentence of the document. However, there is no standard query language despite the wide range of features provided by commercial systems.

On the other hand, traditional information retrieval (IR) systems have a formal foundation based on early library applications and other well-structured textual databases [29, 30]. However, those foundations are not suitable for texts without a fixed structure or no structure at all because they assume that a text is based on words and documents. In hypermedia or genetic databases those assumptions are not valid.

Querying is just one part of the semantic process. The another part is to select the document that you are looking for. Many researchers in IR have pointed out the problems concerned with the user's understanding of the system due to poor interfaces (for example, see [9, 17]). We present yet another visual interface to browse over queries, documents and text structure and contents. This interface is based on an hybrid model for textual databases [3, 4], based on the classic IR model and the model used by the PAT text searching system [11, 28], which sees the text as a sequence of characters and no predefined text structure. This model is flexible, extensible, powerful and rather general.

Although recently there has been several papers dealing with visualization of databases, our approach captures previous work with several new ideas. We present a set of visual tools that allow query manipulation and document analysis and browsing. The main issue is how to visualize large sets of documents. The same ideas can be applied to large sets of files or network addresses, present as results in many operating systems or Internet tools. A graphical query language based on the card paradigm has been already developed and included in a commercial product [1], but it is not included here, and is presented in a forthcoming paper [6].

We first summarize previous work on the topic, followed by the description of the text database model used. The main section present our ideas for a visual interface to handle queries, sets of documents and their contents, as well as a visual analysis tool. The ideas presented here are the synthesis of the author experience in several software projects related to full text retrieval systems and user interfaces [13, 2, 25, 1]. Although some of the ideas presented here are not new, we believe that the relevance of the paper is to look at them in an integrated manner and within the context of large full-text databases.

[*] This work was partially supported by Grant 1940269 from Fondecyt. Contact address during 1996 while in a sabbatical is Dept. LSI, Univ. Politécnica de Catalunya, Barcelona, Spain. E-mail: rbaeza@goliat.upc.es.

AVI '96, Gubbio Italy
© 1996 ACM 0-89791-834-7/96/05..$3.50

2 Previous Work

Most visual representations focus on some specific aspects. In text retrieval we can distinguish visualizations for a single document, several documents or queries. Most of the time only one of those elements is visualized. In the last years, several visual metaphors have been designed. Below we present some of them.

A general user interface framework and interaction is presented in the InfoGrid [23]. In [18] and [8] the semantic organization is addressed, which in our case would be just one view of the document space. The VIBE system [22] also focuses on the document space, but it is based on user given points of interest of the query (using weighted attributes). Another metaphor for the document space based on inter-particle forces, as VIBE, is proposed in [7]. In [32] the document space is abstracted from a Venn diagram to an iconic display called InfoCrystal. One advantage of this scheme is that is also a visual query language. Visual tools in three-dimensions to handle the document space are presented in LyberWorld [15]. A more integrated visual scheme is given in [12], which is based on the query structure. Visualization of occurrence frequency of terms in different text segments of a document is presented in [14]. Specific visualizations applied to text are presented in [24, 10]. More powerful visualizations are available, including three-dimensional visualization, but they need fast hardware [26].

3 Text and Query Model

Let *text* be the data to be searched. This data may be stored in one or more files. It is not necessarily textual data, just a sequence of characters. When the text is large, fast searching is provided by building an *index* of the text, which is used in subsequent searches. We assume that a searching engine of such kind is available. This engine implements a given query language which is the interface to ours. To improve retrieval capabilities and to simplify posing the query, the index is built over a *normalized* text. Text normalization is achieved by processing the original text using a user-defined set of transformations which include stop words, synonyms, character suppression, character translation, etc.

Depending on the retrieval needs, the user must define which positions of the text will be indexed. Every position that must be indexed is called an *index point*. The index points are specified over the normalized text. In addition to the index points, we may specify pieces of the text that will not be indexed (for example, if a text contains images or non-indexable data).

The search for a word or prefix returns all pieces of text (matches or occurrences) matching it in the index (that is, text positions that are index points). Each occurrence is an index point plus its length used to highlight the piece of text that matches the query. Note that only pieces of text starting at an index point can be retrieved by a query, and thus only those can be occurrences. The answer and the original text are used by the user interface to display the actual matches.

Optionally, a text may have a structure. This structure can and must be defined by the application programmer or the user. The text can be divided into *documents*. The text itself may be considered as one document. Each document may be divided into *fields*.

We formally define a *query* as an operation over the text that returns two types of objects:

1. a set of occurrences identified by their position in the text and the length and scope of the text that matches, and
2. a set of files (if the text has no structure) or a set of logical documents (if the user defines a structure).

When there is no structure, a query returns at least the name of the file that stores the text. The first consequence of this definition is that boolean operators are necessary for two different scopes: sequences of symbols (occurrences) and the physical or user-defined structure. Operations to give the subset of occurrences that belong to a given document are also necessary. For more details on the text model and query language see [3, 4]. A comparison between different models that allows queries also in the text structure, including the model used in this paper, is given in [5].

A formal description of the result of a query Q is a set of documents (or files) $D = \{d_1, ..., d_n\}$ and a set of text positions (matches) $M = \{m_1, ..., m_p\}$ with $p \geq n$. Each document has a set of attributes $A = \{a_1, ..., a_\ell\}$. We assume that document attributes can be ordered and have a minimum and maximum value. Possible attributes are: logical or physical position (identifier), temporal values (creation or last modification date), document size, etc.

Our visual interface models the browsing and filtering process with these three elements: queries, documents (possibly with their structure) and text positions. Visualization of queries allows to follow the history of the filtering process. Visualization of the result of each query (documents and text positions) allows to understand the result and facilitates the semantic process done by the user.

4 Visual Browsing

We associate to each element (queries, documents and text positions) one or more views which are related to at least one measure (numerical attribute), which is selected by the user. An example of the user interface is given in Figure 1. The screen is divided vertically into the three elements from right to left: queries, documents, and occurrences. We describe them in this order and after we give some examples of their use as filtering tools.

The user interaction is very simple. The user chooses the desired view for each element. By clicking in a query, the document and text view change to that query answer. By clicking in a document the text view shows the content of it.

4.1 Query Visualization

Queries are arranged from top to bottom (the most recent on top), with the current query highlighted. In the example we show three different views of each query. The pie view is based on the measure given by the number of documents of the database selected. Other views are obtained by pointing to the query. We show two of them as smaller windows on Figure 2(c). The one above, shows the distribution of occurrences within terms of the query (query map) using boxes of different sizes. This view is useful to know what terms are the best filters in the given query. Below, we show the distribution of documents selected on the database logical space, that is, the underlying document identifier mapping (universe map). This view can show if there is any logical locality of reference associated with the query.

More views are possibly, but depend on the specific query language. Also, a specific view (concept) may have more than one meaningful representation. However, the examples above show what is common to all text retrieval queries:

- Query answer size with respect to the universe.
- Query elements "weight".
- Locality of reference in the answer.

4.2 Document Visualization

The central portion of the screen shows the document space selected by the query. There are several measures that can be associated with these documents. For example, number of occurrences in a document or any document attribute. In the example we have two, which generates two different views. The top view is the document space viewed as a non-uniform grid displayed as a fish-eye[1], where the grid focus changes as a pointer device moves (say a mouse) and the current document is selected with a mouse button. The fish-eye concept is only used if the number of documents does not fit within the resolution and size of the current document window. In this view each document is represented in a different gray scale or color, according to a document attribute. In the figure example we use the number of occurrences of the query terms in it. The current document is highlighted, and shows the current value of its attribute. In fact, this is an example of the more general concept of assigning a visual mapping to document attributes. The following visual mappings are possible, and are the buttons labeled on the bottom of the document space in Figure 1.

- Order: the order of the documents can be changed by a given attribute (provided that it has a total order). The order is taken from top to bottom, and left to right.
- Color: the color or gray scale is associated with the attribute. The number of color or gray scales can be automatically generated to a given number of levels (default or specified by the user).
- Size: the size of the square (vertically) can also be variable and associated with an attribute. If we want to use both dimensions, a possible way to do this can be based on tree-maps [31], which depicts trees as squares of different size, but it is less efficient.

For example, with the order and color buttons we can sort the documents by number of occurrences (this might be useful if we know in advance that the query appears many –dark color– or few –white color– times on the document).

The previous view does not use the fact that the screen has two dimensions. Another view uses two dimensions, and is based on the occurrences and the document structure (if exists). The vertical axis (fish-eyed) represents documents and the horizontal axis their structure (see Figure 2(b)). Every field is displayed according to its relative size inside the document and has a bar which depends on a given measure for each field ("F" button). Again, the document space can be sorted with a different attribute on an specific field instead of document order.

The examples above show what is generic to documents. Each document typically has several attributes as well as values that depend on the query, which can be used with a visual mapping. Default values would be logical identifier for the order, number of occurrences for the color, and uniform size. Some of these attributes can be applied to the structure, for example size of an specific field or density of occurrences.

4.3 Text Visualization

The measure associated with occurrences is their position with respect to the whole database or to a given document. The later is used in the example. One possible view, as shown, is a window of the document with

[1] This technique has been used successfully in other applications to show large objects, for example graphs [20].

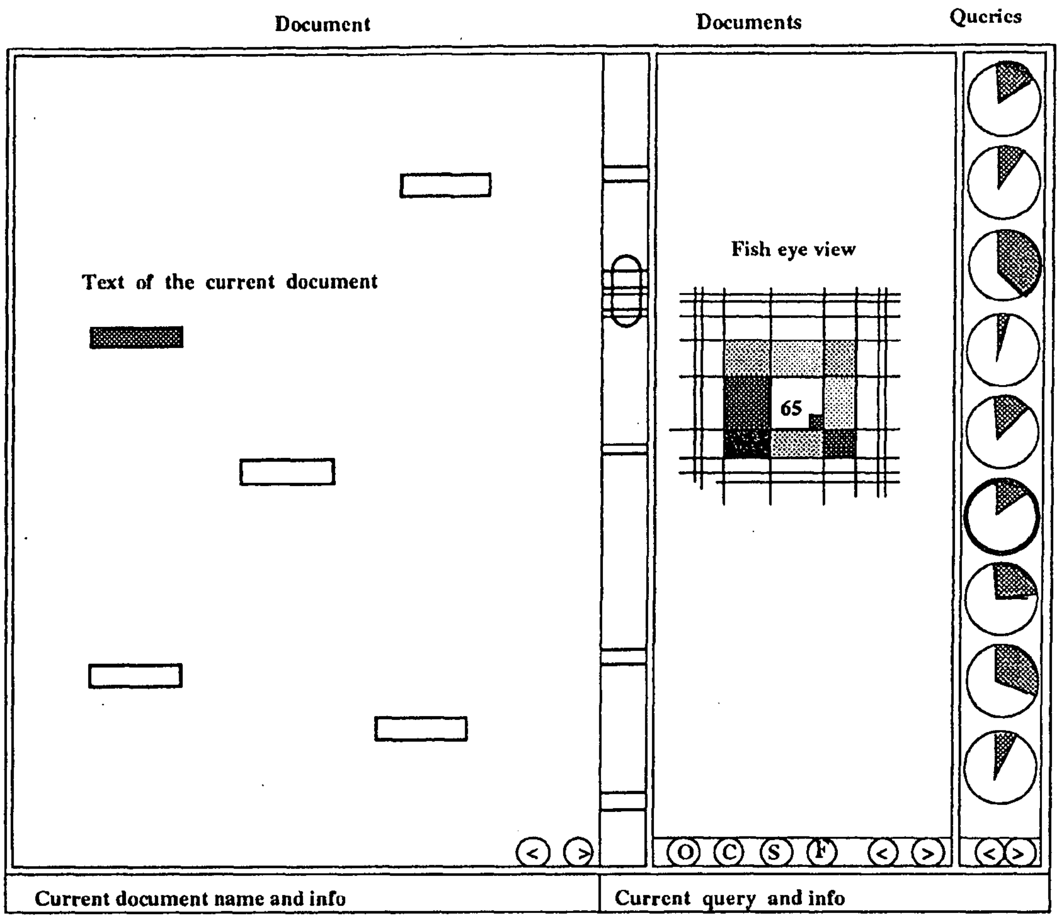

Figure 1: Visual user interface.

the text positions highlighted. The window has an augmented scroll-bar (similar to [21] but in a different context) which has marks where the text positions appear in the document. These marks may depend on actual positions, density, etc; and they can have variable width and/or length.

The scroll-bar can be viewed as a complete compact view of the text. Clicking in any position of the scroll-bar focus the text view in that part of the document. If the document is very large, the scroll-bar can also be fish-eyed with indications of where the nearby occurrences are. The same ideas can be generalized to provide different granularities of the text view:

- The text itself is fish-eyed zooming where the query occurs given some adjacent lines to understand the context (see Figure 2(a) left). The number of lines can be modified by the user.
- Only the text layout is given, in multiple columns, as in [10] (see Figure 2(b) right). Colored (darker) parts indicate lines where the query occurs.

5 Answer Filtering and Selection

In this section we present a more elaborate metaphor for manipulating and filtering an answer given by a set of documents. Figure 3 shows an instance of the visual analysis tool that we propose for advanced users. We use a "library" or "bookpile" analogy depending if the tool is used horizontally or vertically, because both are possible. The "pile" metaphor has been used before, but in a different way and for different purposes (for example see [27, 16]). Each document (seen as a book) is represented as a rectangle with a particular color, height, width and position into the set. Each one of this graphical attributes, including the order of the list, can be mapped to a document attribute (occurrence density, size, date, etc). In the example, the order and the color are mapped to the same attribute (for example, the creation date). These mappings allow to study different correlations of attributes on the document set, helping the user to select the desired documents. A select button allows to choose a document subset by using the mouse (the wide border rectangle in the example).

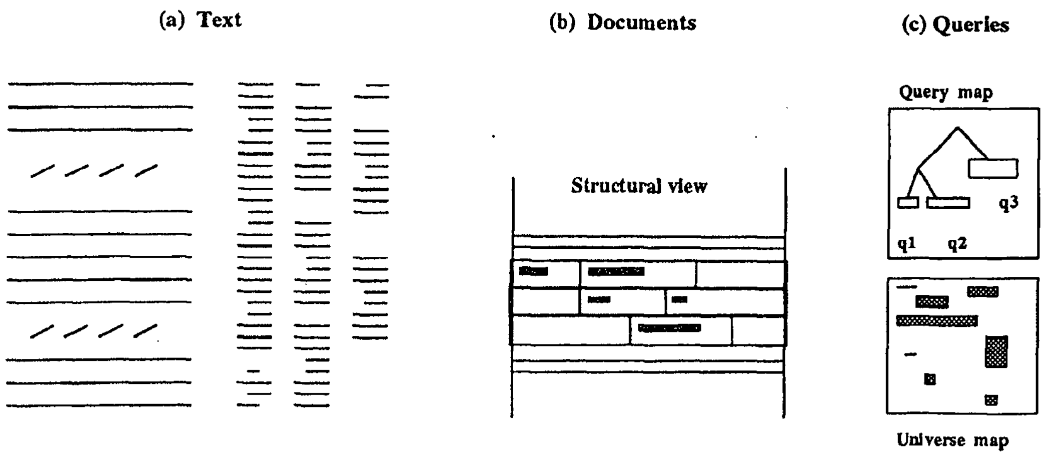

Figure 2: Other views.

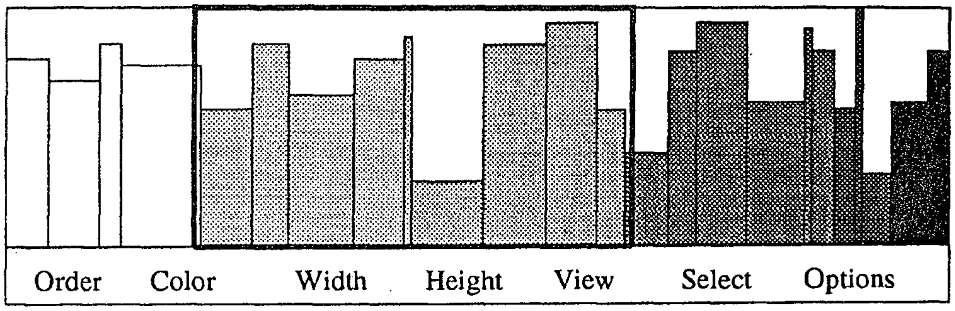

Figure 3: Analyzing and selecting a document set.

The mapping of the attributes is selected by the menu buttons below the book list. The mapping can be done linearly or logarithmicly (in the case of attributes with large scales). The way the books are seen can also be changed. The set of documents can be forced to fit into the window (as in the example), presented using a predefined choice of maximal/minimal widths and heights (using a scroll-bar if bigger than the window). Another view is a fish-eye representation for large sets, focusing where the user wants (by clicking with the mouse the appropriate sector).

6 Concluding Remarks

The visual tool interface presented here should be considered just as different views for the same data. The main idea is that the user chooses the view which is most suitable for his/her semantic process or filtering needs. That depends on what the user has in mind, his/her knowledge of the database, and the type of application.

Some of the views presented for the query level can be adapted for a visual query language or to allow direct manipulation of the query from the visual representation. They can also be used to associate weights to each element of the query. At the document or text level, a possible extension is to allow user defined markers chosen from a set of standard marks which may convey different type of annotations and/or degree of interest.

We are currently working in a prototype of this interface using HotJava, which includes the views given in the example, as well as other views and tools. In particular, visual tools are a must considering the amount of text data currently available in Internet. We want to extend this interface to related tasks on distributed and collaborative text databases. We also want to include more elaborate queries which will include also the text structure (for example in SGML), using the query language that we proposed in [19].

The visual representations presented allow to perform several different correlations between the documents of the result. Those correlations help the semantic analysis of the user. For example, we could correlate up to four attributes of documents by choosing the order, color and size of the document space representation. So, we can see locality related to date with respect to number of occurrences or similar relations, depending on partial information already known by the user but not easy to represent in the query language. That is a key issue:

some knowledge may difficult to formalize, but easier to visualize.

The usefulness of the interface presented must have several usability tests to obtain user feed-back and to improve its design. The interface is not only useful for classical text retrieval systems, but can be applied to many other tools which have to represent data structured in two levels. Some examples follow:

- Large file systems: for example a file specification as "*.c" or a grep query can be displayed similarly. Documents in this case are files.
- Network addresses plus files: for example global searches using Archie (sources for public software) or in WWW (URLs).

Acknowledgements

We wish to acknowledge the helpful comments of Mariano Consens, Nancy Hitschfeld, Pablo Palma, Roxana Sagues and the unknown referees.

REFERENCES

[1] ARS INNOVANDI. *Search City*, 1991. User's Manual.

[2] BAEZA-YATES, R. *Efficient Text Searc'.ing.* PhD thesis, Dept. of Computer Science, University of Waterloo, May 1989. Also as Research Report CS-89-17.

[3] BAEZA-YATES, R. Modeling, browsing and querying large text databases. Tech. Rep. DCC-94-2, Dept. of Computer Science, Univ. of Chile, 1994.

[4] BAEZA-YATES, R. An extended model for full-text databases. *Journal of Brazilian CS Society* (1996). to appear.

[5] BAEZA-YATES, R., AND NAVARRO, G. Integrating contents and structure in text retrieval. *ACM SIGMOD Record* (May 1996).

[6] BAEZA-YATES, R., AND PALMA, P. A graphical query language for full text retrieval. Dept. of Computer Science, Univ. of Chile and Ars Innovandi (in preparation), 1995.

[7] CHALMERS, M., AND CHITSON, P. BEAD: Exploration in information visualization. In *ACM SIGIR'92* (1992).

[8] CHANG, S.-K. Visual reasoning for IR from very large databases. In *IEEE Workshop on Visual Languages* (1989), pp. 1-6.

[9] CROUCH, D. The visual display of information in an IR environment. In *ACM SIGIR* (1989), pp. 58-67.

[10] EICK, S. Graphically displaying text. *Journal of Computational and Graphical Statistics 3*, 2 (1994), 127-142.

[11] FAWCETT, H. *A Text Searching System: PAT 3.3, User's Guide*. Centre for the New Oxford English Dictionary, University of Waterloo, 1989.

[12] FOWLER, R., FOWLER, W., AND WILSON, B. Integrating query, thesaurus, and documents through a common visual representation. In *ACM SIGIR* (1991), pp. 142-151.

[13] GONNET, G. Examples of PAT applied to the Oxford English Dictionary. Tech. Rep. OED-87-02, Centre for the New OED., University of Waterloo, 1987.

[14] HEARST, M. Tilebars: Visualization of term distribution information in full text information access. In *ACM SIGCHI* (Denver, CO, May 1995).

[15] HEMMJE, M., KUNKEL, C., AND WILLET, A. Lyberworld - a visualization user interface supporting text retrieval. In *17th ACM SIGIR* (Dublin, Jul 1994).

[16] KIM, H., AND KORFHAGE, R. Bird: Browsing interface for the retrieval of documents. In *IEEE Symp. on Visual Languages* (St. Louis, Missouri, Oct 1994), pp. 176-177.

[17] KORFHAGE, R. To see or not to see: is that the query? In *14th ACM SIGIR* (Chicago, 1991).

[18] LIN, X., SOERGEI, D., AND MARCHIONINI, G. A self-organizing map for IR. In *14th ACM SIGIR* (1991), pp. 262-269.

[19] NAVARRO, G., AND BAEZA-YATES, R. A language for queries on structure and contents of textual databases. In *18th ACM Conference on Research and Development in Information Retrieval (SIGIR'95)* (Seattle, WA, USA, July 1995).

[20] NOIK, E. Exploring large hyperdocuments: Fisheye views of nested networks. In *ACM Hypertext '93* (1993), pp. 192-205.

[21] OLSEN, D. Bookmarks: An enhanced scroll bar. *ACM Trans. on Computer Graphics 11*, 3 (1992), 291-295.

[22] OLSEN, K., KORFHAGE, R., SOCHATS, K., SPRING, M., AND WILLIAMS, J. Visualization of a document collection: The VIBE system. *Information Processing and Management 29*, 1 (1993), 69-81.

[23] Rao, R., Card, S., Jellinek, D., Mackinlay, J., and Robertson, G. The information grid: A framework for information retrieval and retrieval-centered applications. In *UIST'92* (Monterey, California, 1992), pp. 23–32.

[24] Raymond, D. Visualizing texts. In *Making Sense of Words: 9th Annual Conference of the UW Centre for the New Oxford English Dictionary* (Oxford, England, Sept 1993), pp. 20–33.

[25] Raymond, D., and Fawcett, H. Playing detective with full text searching software. Tech. Rep. OED-90-03, Centre for the New OED., University of Waterloo, 1990.

[26] Robertson, G., Card, S., and Mackinlay, J. Information visualization using 3d interactive animation. *Comm. of the ACM 36*, 4 (1993), 56–71.

[27] Rose, D., Mander, R., Oren, T., Ponceleón, D., Salomon, G., and Wong, Y. Content awareness in a file system intergace: Implementing the 'pile' metaphor for organizing information. Tech. rep., Apple Computer, Inc., 1993.

[28] Salminen, A., and Tompa, F. Pat expressions: an algebra for text search. In *Second Conference on Computational Lexicography* (Budapest, Oct 1992).

[29] Salton, G., and McGill, M. *Introduction to Modern Information Retrieval*. McGraw-Hill, New York, 1983.

[30] Salton, G., and McGill, M. *Automatic Text Processing*. Addison-Wesley, Mass, 1989.

[31] Shneiderman, B. Tree visualization with tree-maps: 2-d space-filling approach. *ACM Trans. on Computer Graphics 11*, 1 (1992), 92–99.

[32] Spoerri, A. Infocrystal: A visual tool for information retrieval and management. In *Information and Knowledge Management'93* (Washington D.C., 1993).

Citation

AVI >archive
Proceedings of the workshop on Advanced visual interfaces >toc
1996 , Gubbio, Italy

SESSION: Interfaces to databases >toc

Visualization of large answers in text databases

Author
Ricardo Baeza-Yates Univ. of Chile, Blanco Encalada 2120, Santiago, Chile

Sponsor
SIGMULTIMEDIA : ACM Special Interest Group on Multimedia

Publisher
ACM Press New York, NY, USA

Pages: 101 - 107 Series-Proceeding-Section-Article
Year of Publication: 1996
ISBN:0-89791-834-7

doi> http://doi.acm.org/10.1145/948449.948464 (Use this link to Bookmark this page)

> full text > abstract > references > index terms > peer to peer

> Discuss > Similar > Review this Article Save to Binder

> BibTex Format

↑ **FULL TEXT:** **Access Rules**

pdf 737 KB

↑ **ABSTRACT**

Current user interfaces of full text retrieval systems do not help in the process of filtering the result of a query, usually very large. We address this problem and we propose a visual interface to handle the result of a query, based on a hybrid model for text. This graphical user interface provides several visual representations of the answer and its elements (queries, documents, and text), easing the analysis and the filtering process.

↑ REFERENCES

Note: OCR errors may be found in this Reference List extracted from the full text article. ACM has opted to expose the complete List rather than only correct and linked references.

1 ARS INNOVANDI. Search City, 1991. User's Manual.

2 BAEZA-YATES, R. Efficient Text Searching. PhD thesis, Dept. of Computer Science, University of Waterloo, May 1989. Also as Research Report CS-89-17.

3 BAEZA-YATES, R. Modeling, browsing and querying large text databases. Tech. Rep. DCC-94-2, Dept. of Computer Science, Univ. of Chile, 1994.

4 BAEZA-YATES, R. An extended model for full-text databases. Journal of Brazilian CS Society (1996). to appear.

5 Ricardo Baeza-Yates , Gonzalo Navarro, Integrating contents and structure in text retrieval, ACM SIGMOD Record, v.25 n.1, p.67-79, March 1996

6 BAEZA-YATES, R., AND PALMA, P. A graphical query language for full text retrieval. Dept. of Computer Science, Univ. of Chile and Ars Innovandi (in preparation), 1995.

7 Matthew Chalmers , Paul Chitson, Bead: explorations in information visualization, Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval, p.330-337, June 21-24, 1992, Copenhagen, Denmark

8 CHANG, S.-K. Visual reasoning for IR from very large databases. In IEEE Workshop on Visual Languages (1989), pp. 1--6.

9 CROUCH, D. The visual display of information in an IR environment. In ACM SIGIR (1989), pp. 58--67.

10 EICK, S. Graphically displaying text. Journal of Computational and Graphical Statistics 3, 2 (1994), 127--142.

11 FAWCETT, H. A. Text Searching System: PAT 3.3, User's Guide. Centre for the New Oxford English Dictionary, University of Waterloo, 1989.

12 Richard H. Fowler , Wendy A. L. Fowler , Bradley A. Wilson, Integrating query thesaurus, and documents through a common visual representation, Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval, p.142-151, October 13-16, 1991, Chicago, Illinois, United States

13 GONNET, G. Examples of PAT applied to the Oxford English Dictionary. Tech. Rep. OED-87-02, Centre for the New OED., University of Waterloo, 1987.

14 Marti A. Hearst, TileBars: visualization of term distribution information in full text information access, Proceedings of the SIGCHI conference on Human factors in computing systems, p.59-66, May 07-11, 1995, Denver, Colorado, United States

15 Matthias Hemmje , Clemens Kunkel , Alexander Willett, LyberWorld—a visualization user interface supporting fulltext retrieval, Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval, p.249-259, July 03-06, 1994,

Dublin, Ireland

16 KIM, H., AND KORFHAGE, R. Bird: Browsing interface for the retrieval of documents. In IEEE Symp. on Visual Languages (St. Louis, Missouri, Oct 1994), pp. 176--177.

17 Robert R. Korfhage, To see, or not to see— is That the query?, Proceedings of the 14th annual international ACM SIGIR conference on Research and development in information retrieval, p.134-141, October 13-16, 1991, Chicago, Illinois, United States

18 LIN, X., SOERGEI, D., AND MARCHIONINI, G. A self-organizing map for IR. In 14th ACM SIGIR (1991), pp. 262--269.

19 Gonzalo Navarro , Ricardo Baeza-Yates, A language for queries on structure and contents of textual databases, Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, p.93-101, July 09-13, 1995, Seattle, Washington, United States

20 Emanuel G. Noik, Exploring large hyperdocuments: fisheye views of nested networks, Proceedings of the fifth ACM conference on Hypertext, p.192-205, November 14-18, 1993, Seattle, Washington, United States

21 Dan R. Olsen, Jr., The interaction technique notebook: Bookmarks: an enhanced scroll bar, ACM Transactions on Graphics (TOG), v.11 n.3, p.291-295, July 1992

22 Kai A. Olsen , Robert R. Korfhage , Kenneth M. Sochats , Michael B. Spring , James G. Williams, Visualization of a document collection: the vibe system, Information Processing and Management: an International Journal, v.29 n.1, p.69-81, Jan.-Feb. 1993

23 Ramana Rao , Stuart K. Card , Herbert D. Jellinek , Jock D. Mackinlay , George G. Robertson, The information grid: a framework for information retrieval and retrieval-centered applications, Proceedings of the 5th annual ACM symposium on User interface software and technology, p.23-32, November 15-18, 1992, Monteray, California, United States

24 RAYMOND, D. Visualizing texts. In Making Sense of Words: 9th Annual Conference of the UW Centre for the New Oxford English Dictionary (Oxford, England, Sept 1993), pp. 20--33.

25 RAYMOND, D., AND FAWCETT, H. Playing detective with full text searching software. Tech. Rep. OED-90-03, Centre for the New OED., University of Waterloo, 1990.

26 George G. Robertson , Stuart K. Card , Jack D. Mackinlay, Information visualization using 3D interactive animation, Communications of the ACM, v.36 n.4, p.57-71, April 1993

27 ROSE, D., MANDER, R., OREN, T., PONCELEÓN, D., SALOMON, G., AND WONG, Y. Content awareness in a file system intergace: Implementing the 'pile' metaphor for organizing information. Tech. rep., Apple Computer, Inc., 1993.

28 SALMINEN, A., AND TOMPA, F. Pat expressions: an algebra for text search. In Second Conference on Computational Lexicography (Budapest, Oct 1992).

29 Gerard Salton , Michael J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, Inc., New York, NY, 1986

30 Gerald Salton, Automatic text processing, Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 1988

31 Ben Shneiderman, Tree visualization with tree-maps: 2-d space-filling approach, ACM Transactions on Graphics (TOG), v.11 n.1, p.92-99, Jan. 1992

32 Anselm Spoerri, InfoCrystal: a visual tool for information retrieval & management, Proceedings of the second international conference on Information and knowledge management, p.11-20, November 01-05, 1993, Washington, D.C., United States

↑ INDEX TERMS

Keywords:
set visualization, visual analysis, visual browsing, visual query languages, visual representations, visual text database, visual tools

↑ Peer to Peer - Readers of this Article have also read:

◆ Routing with guaranteed delivery in ad hoc wireless networks
Proceedings of the 3rd international workshop on Discrete algorithms and methods for mobile computing and communications
Prosenjit Bose , Pat Morin , Ivan Stojmenović , Jorge Urrutia

◆ Location-aware query processing in mobile database systems
Proceedings of the 1998 ACM symposium on Applied Computing
Hans-Erich Kottkamp , Olaf Zukunft

◆ Editorial
interactions 8, 5
Steven Pemberton

◆ A measurement-analytic approach for QoS estimation in a network based on the dominant time scale
IEEE/ACM Transactions on Networking (TON) 11, 2
Do Young Eun , Ness B. Shroff

◆ What's happening
interactions 8, 5
Marisa Campbell